
# Common **Meta**data **for** Climate Modeling Digital Repositories




Allyn Treshansky

GIP Kick Off Meeting, Princeton 5-6 November 2009

# Outline

- why care about METAFOR?
- who is METAFOR?
- what has (*and hasn't*) METAFOR done?
  - The CIM
  - The Query Tool(s)
  - The Questionnaire

# why?

- finding data is hard
- understanding data is harder (esp. for non-experts)
- comparing models is confusing
- documentation is “patchy”
- different modeling centres are rather idiosyncratic

# why?

- For example...
  - CMIP5 will involve more than 20 different models from at least 20 institutions running dozens of simulations corresponding to dozens of experiments
  - The volume of data will be petabytes; even the core data will be of order ¾ petabyte
  - *How do users decide what to look at?*
  - *How do they discriminate between datasets?*
  - *How do they find useful documentation?*

# who?

- PRISM [2001 – 2004]

  *common model infrastructures*

- PRISM Sustained Initiative (PSI) [2004→]

  *useful stuff from the folks who brought you PRISM*

  - METAFOR: *metadata standards, data management*
  - IS-ENES: *coupling & computing issues, modeling environments & integration*

- ENES [to infinity and beyond]

  *coordinate earth system modeling w/in Europe*

# who?

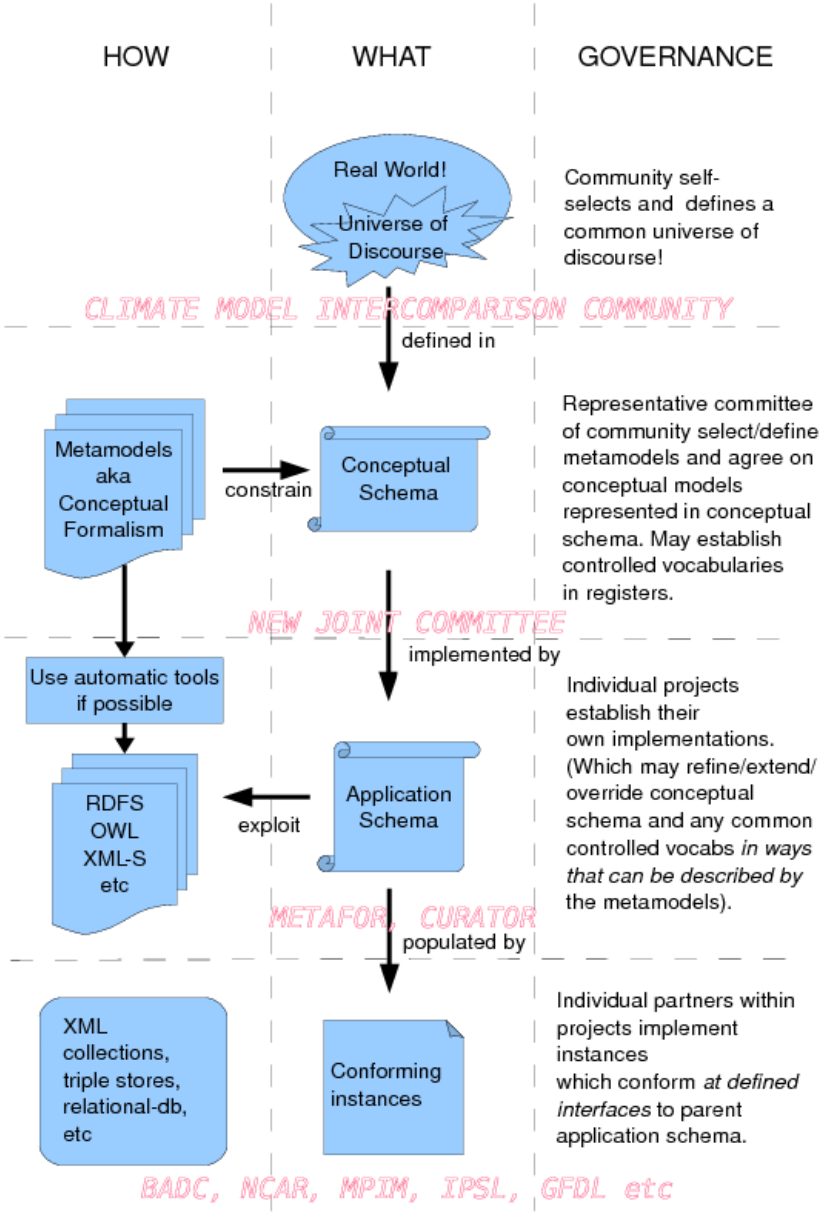
- 3 year project, began March 2008
- EU contribution of 2.2M €
- 11+ partners...

# what?

- another metadata standard (the CIM)
- some clever ways, using metadata, of looking for climate data (the Query Tool)
- a procedure for publishing/archiving metadata (the Questionnaire)

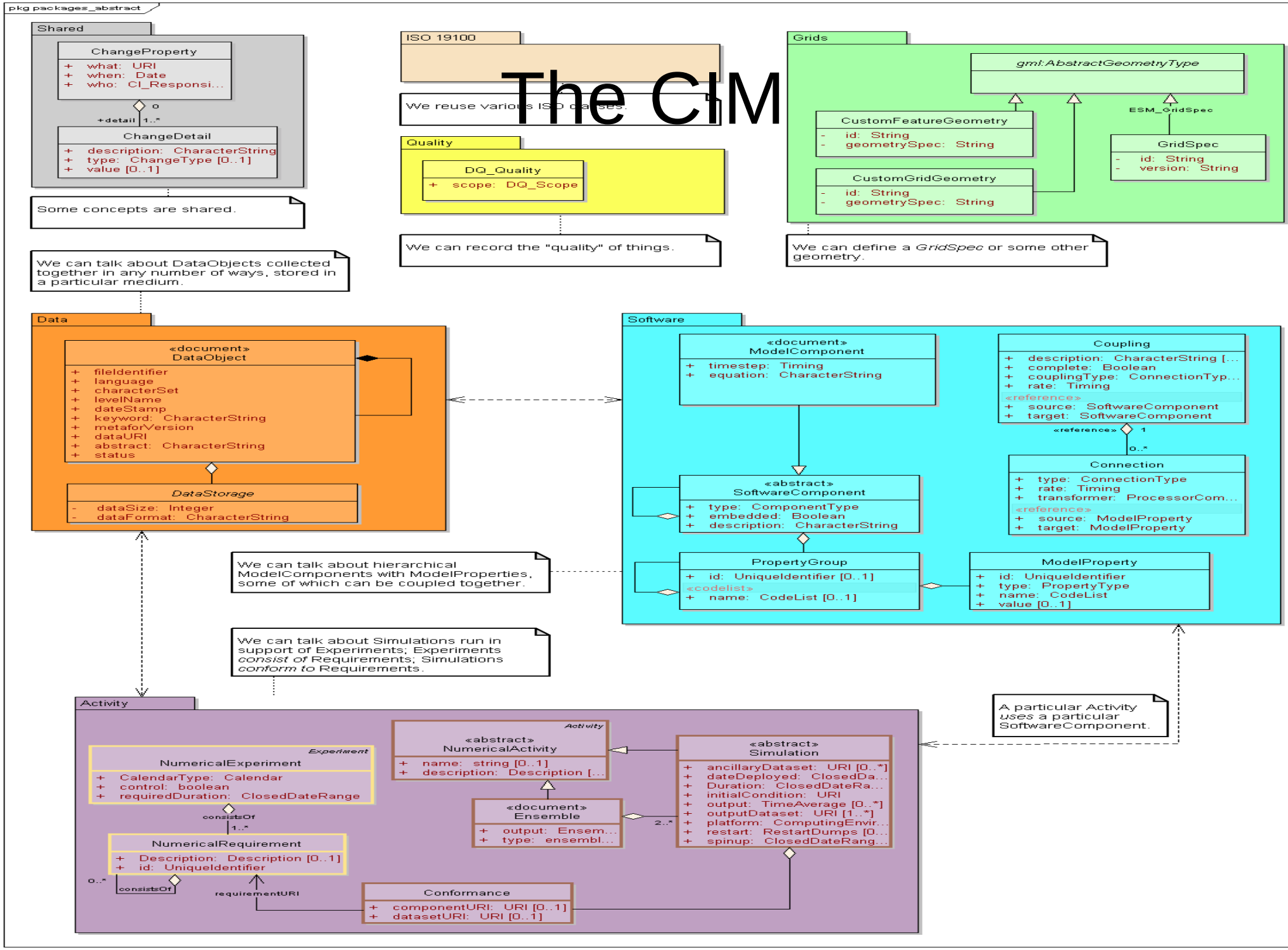
# The CIM

- lots of people talk about climate models and data;
- some people even agree about those things;
- we have a formal way of describing that (UML, *CONCIM*);
- that UML is constrained to follow a particular meta-model...
- ...so that it can be transformed into something usable (*XSD, APPCIM*) for particular users;
- metadata instances conform to an APPCIM



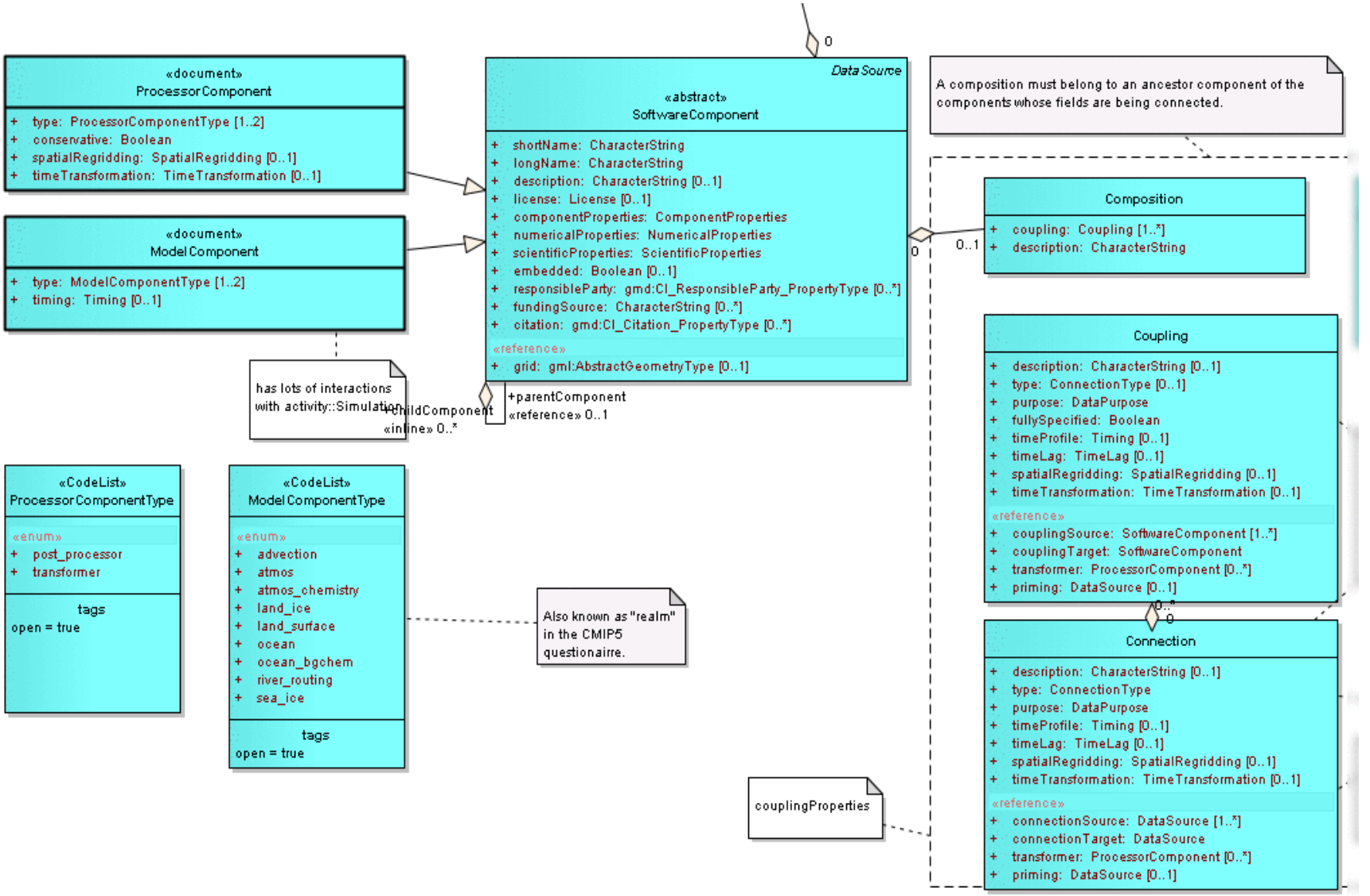
# The CIM

- the CIM builds on existing metadata standards used internationally in climate research (CF, NMM, Curator, FLUME, ISO-standards, etc.)
- the CIM defines a general structure over which a specific Controlled Vocabulary (CV) can be applied
  - a CV consists of the terms (and their relationships) used to build the content of CIM instances.


pkg packages_abstract
Shared
ChangeProperty
+ what: URI
+ when: Date
+ who: CI_Responsi...
0
+detail 1..*
ChangeDetail
+ description: CharacterString
+ type: ChangeType [0..1]
+ value [0..1]
Some concepts are shared.
ISO 19100
We reuse various ISO classes.
Quality
DQ_Quality
+ scope: DQ_Scope
We can record the "quality" of things.
Grids
gml:AbstractGeometryType
CustomFeatureGeometry
- id: String
- geometrySpec: String
CustomGridGeometry
- id: String
- geometrySpec: String
ESM_GridSpec
GridSpec
- id: String
- version: String
We can define a GridSpec or some other geometry.
The CIM
We can talk about DataObjects collected together in any number of ways, stored in a particular medium.
Data
«document»
DataObject
+ fileIdentifier
+ language
+ characterSet
+ levelName
+ dateStamp
+ keyword: CharacterString
+ metaforVersion
+ dataURI
+ abstract: CharacterString
+ status
DataStorage
- dataSize: Integer
- dataFormat: CharacterString
Software
«document»
ModelComponent
+ timestep: Timing
+ equation: CharacterString
Coupling
+ description: CharacterString [...
+ complete: Boolean
+ couplingType: ConnectionTyp...
+ rate: Timing
«reference»
+ source: SoftwareComponent
+ target: SoftwareComponent
«reference» 1
0..*
Connection
+ type: ConnectionType
+ rate: Timing
+ transformer: ProcessorCom...
«reference»
+ source: ModelProperty
+ target: ModelProperty
«abstract»
SoftwareComponent
+ type: ComponentType
+ embedded: Boolean
+ description: CharacterString
PropertyGroup
+ id: UniqueIdentifier [0..1]
«codelist»
+ name: CodeList [0..1]
ModelProperty
+ id: UniqueIdentifier
+ type: PropertyType
+ name: CodeList
+ value [0..1]
We can talk about hierarchical ModelComponents with ModelProperties, some of which can be coupled together.
We can talk about Simulations run in support of Experiments; Experiments consist of Requirements; Simulations conform to Requirements.
Activity
Experiment
NumericalExperiment
+ CalendarType: Calendar
+ control: boolean
+ requiredDuration: ClosedDateRange
consistsOf
1..*
NumericalRequirement
+ Description: Description [0..1]
+ id: UniqueIdentifier
0..*
consistsOf
requirementURI
Activity
«abstract»
NumericalActivity
+ name: string [0..1]
+ description: Description [...
«document»
Ensemble
+ output: Ensem...
+ type: ensembl...
2..*
«abstract»
Simulation
+ ancillaryDataset: URI [0..*]
+ dateDeployed: ClosedDa...
+ Duration: ClosedDateRa...
+ initialCondition: URI
+ output: TimeAverage [0..*]
+ outputDataset: URI [1..*]
+ platform: ComputingEnvir...
+ restart: RestartDumps [0...
+ spinup: ClosedDateRang...
Conformance
+ componentURI: URI [0..1]
+ datasetURI: URI [0..1]
A particular Activity uses a particular SoftwareComponent.

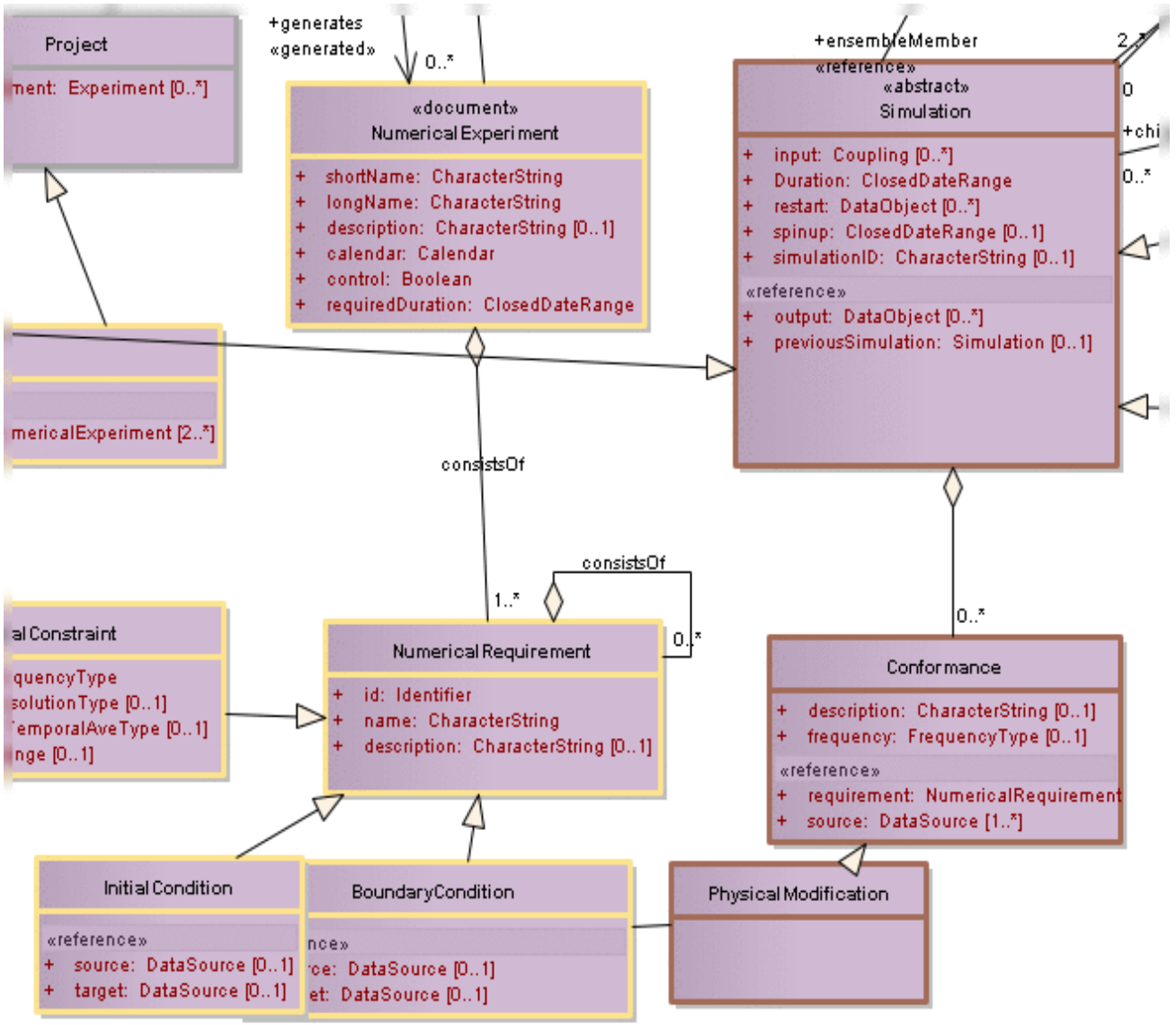
# The CIM

- some interesting bits



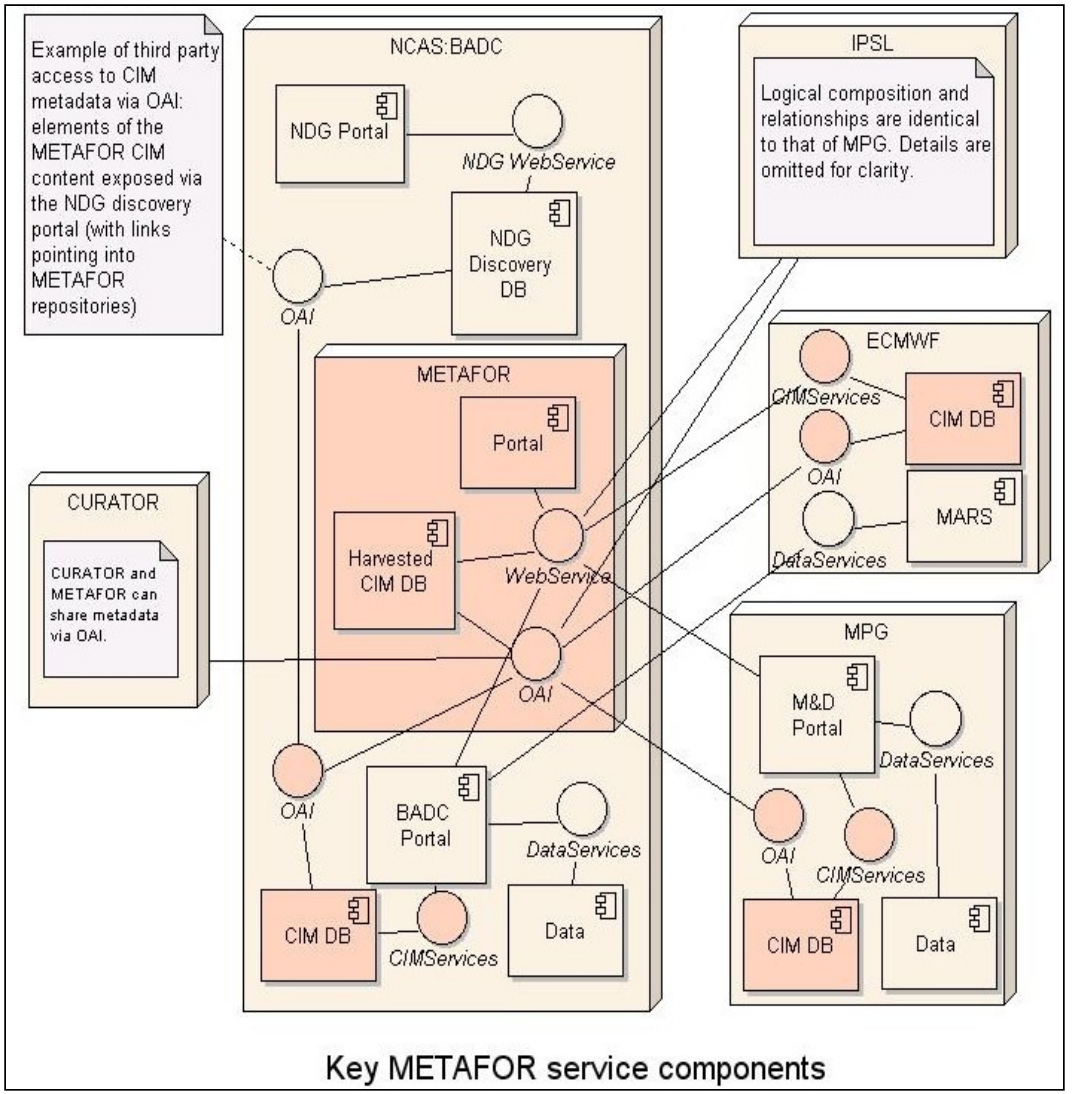
# The CIM

- some other interesting bits



# The Query Tool(s)

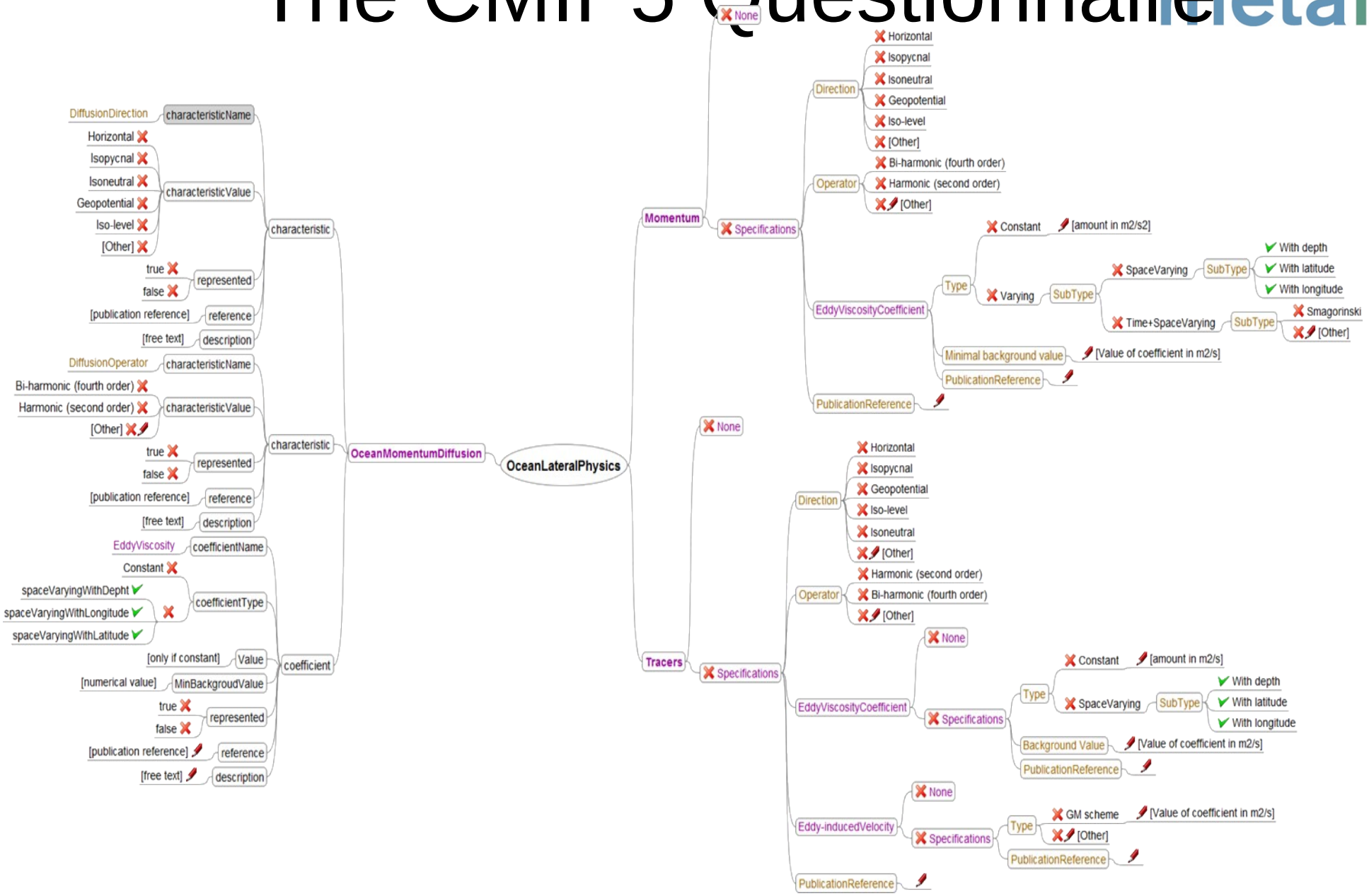
- building a search interface for CIM instances
- building a CIM instance viewer
- building a CIM instance comparer



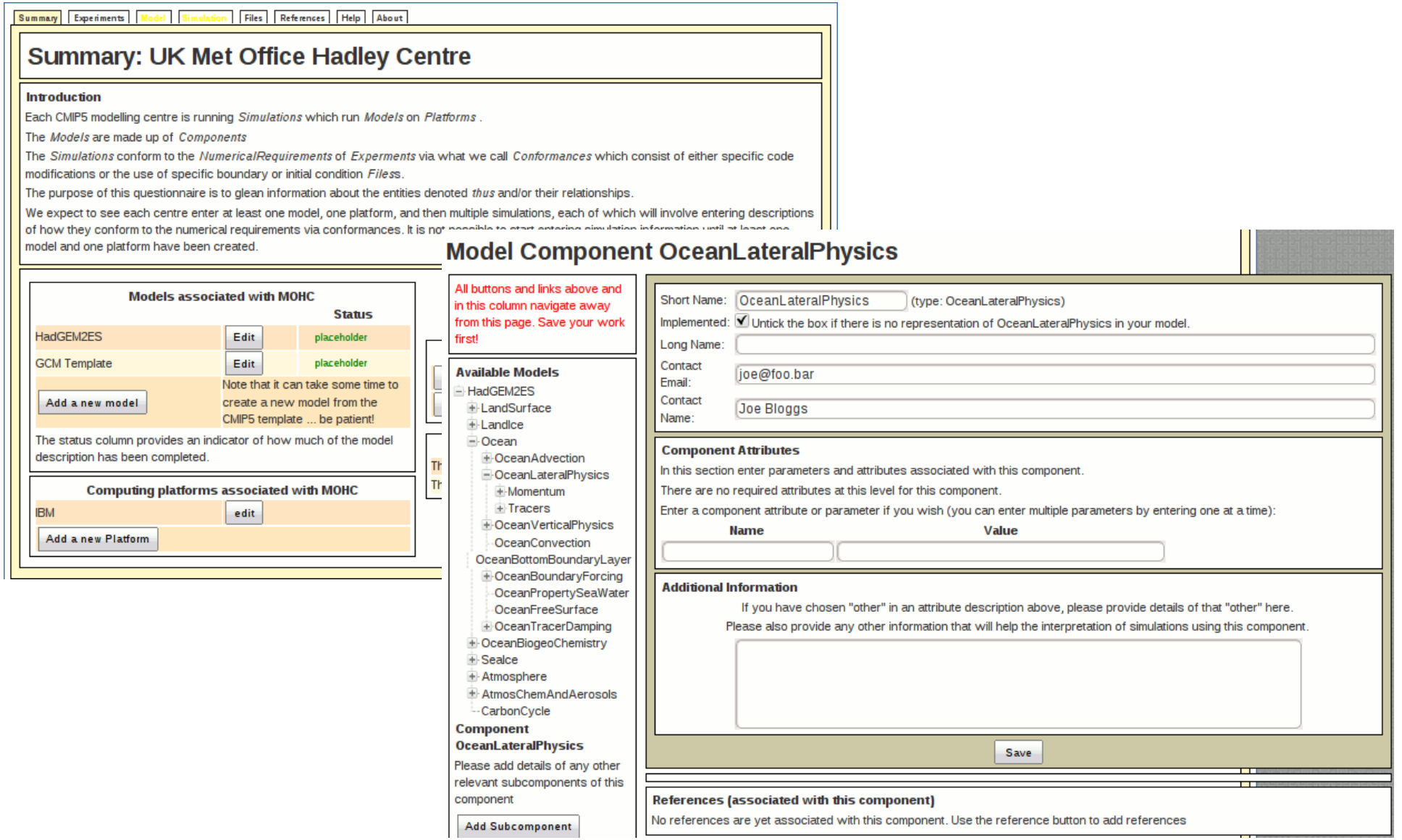
# The CMIP5 Questionnaire

- METAFOR has been tasked by WGCM/CMIP to produce an online tool to capture relevant metadata
- the CMIP5 archives will store ~500TB of data
- we aim to store metadata instances that describe that data, including the software models that generated the data and the experiments for which those models were run
- we aim to collect enough detail to allow users to easily...
  a) browse the archive & find desired datasets
  b) easily differentiate between the “genealogy” (related models & experiments) of datasets

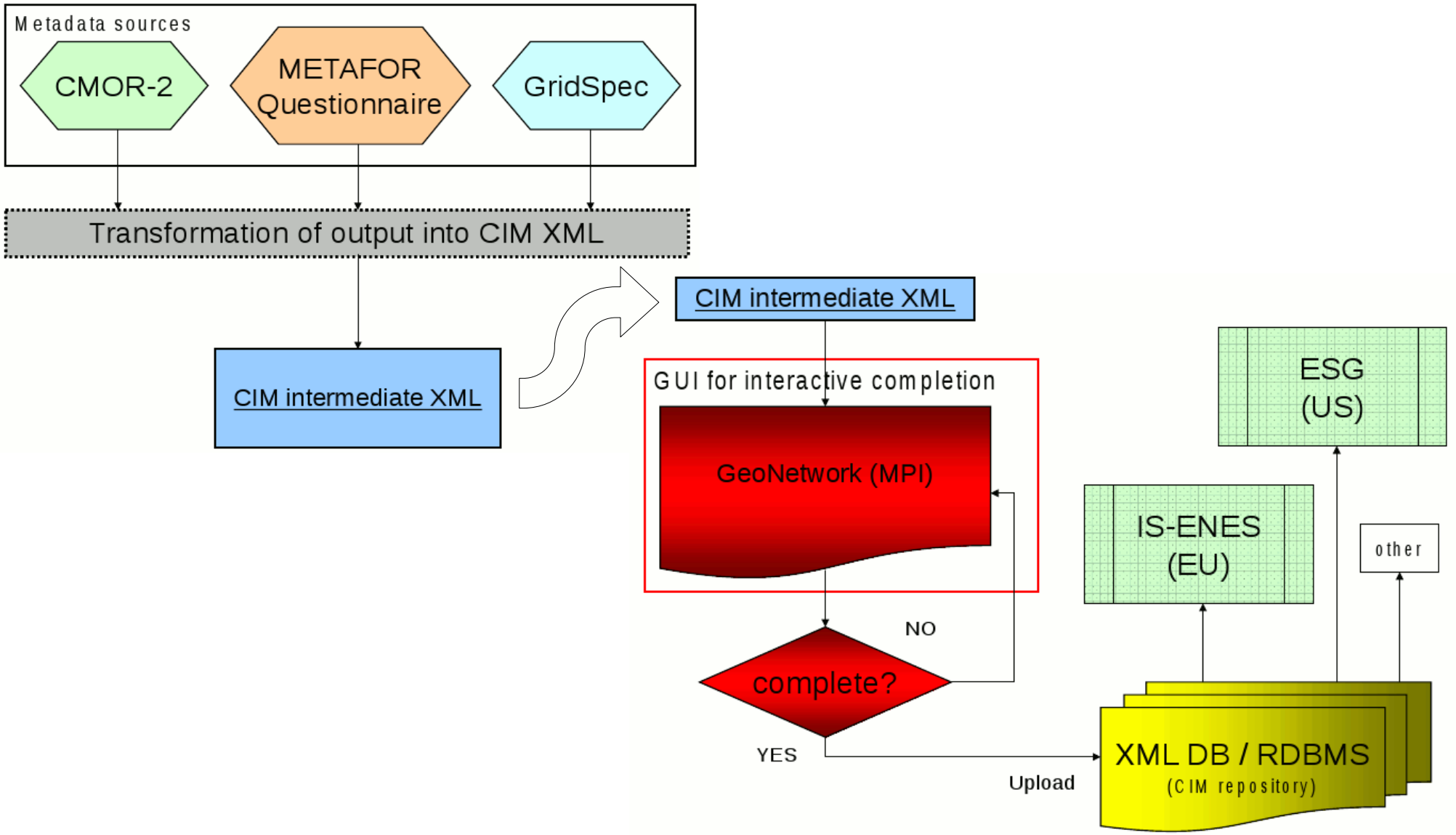
# The CMIP5 Questionnaire



# The CMIP5 Questionnaire



# tweaking things a bit



# The End

- “main” site: http://metaforclimate.eu
- “working” site: http://metaforclimate.eu/trac
- the CIM: http://metaforclimate.eu/trac/browser/CIM
- the questionnaire: http://cmip5q.ceda.ac.uk
- mailing list: metafor@lists.enes.org
- guy-in-charge: Eric.Guilyardi@locean-ipsl.upmc.fr
- me: allynt@coelacanthconsulting.com